

DIE GOOGLE BUCHSUCHE ALS HILFSMITTEL FÜR DIE LEXIKOGRAPHIE¹

von Dominik Brückner

Allgemeines

Google Inc.

Die Firma Google Inc. wurde 1998 von Sergey Brin und Larry Page gegründet. Benannt ist sie nach „Googol“, dem Namen für die Zahl 10^{100} , der 1938 von Edward Kasner etabliert wurde.² Der Börsengang erfolgte im Jahr 2004, am 17. März 2008 betrug der Börsenwert des weltweit fast 20.000 Mitarbeiter beschäftigenden Unternehmens rund 84 Milliarden Euro.

Das bekannteste Produkt der Firma ist die Internetsuchmaschine „Google“, seit einigen Jahren wird jedoch ein weiteres Online-Tool angeboten, das insbesondere in den Textwissenschaften zunehmend Anwendung findet, die Google-Buchsuche (Google Booksearch).

Die Google Buchsuche

Die Buchsuche ist ein kostenloses Online-Tool, das eine Suche nach Texten in vollständig digitalisierten Büchern ermöglicht. Derzeit (Stand: Juli 2009) ist nur eine Betaversion öffentlich verfügbar. Googles Ziele

sind ehrgeizig: Geplant ist die vollständige Digitalisierung sämtlicher Bücher dieser Welt, in allen Sprachen (deren Zahl wird von Google mit 430 angegeben) und aus allen Zeiten. Das Nahziel besteht immerhin in der digitalen Aufbereitung von 15 Millionen Büchern bis 2015.

Das Projekt „Google Buchsuche“ begann im Jahr 2002 mit einer Planungs- und Konzeptionsphase. Im darauffolgenden Jahr wurde ein schonendes Hochgeschwindigkeits-Scanverfahren entwickelt. 2004 schloss Google dann die erste Bibliothekspartnerschaft mit der Bodleian Library (der Universitätsbibliothek von Oxford) und warb auf der Frankfurter Buchmesse erste Partner für sein Verlagsprogramm an (die Verlage stellen Google ihre Produkte direkt zur Verfügung und bestimmen die Nutzung durch Google). Seit 2005 werden technische Verbesserungen und neue Funktionen bereitgestellt. Derzeit sind etwa sieben Millionen Bücher in 35 Sprachen digitalisiert und es bestehen Partnerschaften mit 28 Bibliotheken und rund 20.000 Verlagen in 100 Ländern.

Die Details des Scanverfahrens sind weitgehend geheim. Scans, auf denen Finger oder nur halb umgeblätterte Seiten zu sehen sind und die den Eindruck er-

wecken, Google ginge Geschwindigkeit vor Qualität, verschwinden derzeit nach und nach aus dem System.

Partnerprogramm mit Bibliotheken

Google scannt derzeit (Teil-)Bestände ganzer Bibliotheken ein. Man beschränkt sich dabei (noch) auf Großbibliotheken, vor allem in den USA (z. B. die Universitätsbibliotheken von Princeton, Stanford, Harvard, Columbia, Cornell, Virginia, Michigan, Wisconsin-Madison und die New York Public Library). In Europa beteiligen sich noch wenige Bibliotheken an dem Partnerprogramm (darunter die Universitätsbibliotheken Oxford, Lausanne, Gent und Madrid, die Katalonische Nationalbibliothek, die Stadtbibliothek Lyon und die Bayerische Staatsbibliothek in München). In Japan ist die Universitätsbibliothek Keio beteiligt.

Die Bayerische Staatsbibliothek nimmt als erste deutsche Bibliothek seit März 2007 an dem Digitalisierungsprojekt teil. Geplant ist die Bereitstellung von einer Million gemeinfreier³ Bücher in verschiedenen Sprachen. Die Digitalisierung läuft seit Sommer 2008.

Arbeitsmöglichkeiten

Die Kritik an der Google-Buchsuche ist in den letzten Jahren derart laut geworden, dass sie über die einschlägigen Fachkreise hinaus in die breite Öffentlichkeit gedrungen ist. Vor allem Copyrightverletzungen, die „Hegemonie des Englischen“ (J.-N. Jeanneney), „Knebelverträge“ für Partnerbibliotheken und -verlage, die mäßige OCR-Qualität oder die Konkurrenz zu anderen, öffentlich finanzierten Digitalisierungsprojekten werden kritisiert. Trotz alledem findet die Google-Buchsuche in den letzten Jahren zunehmend (wissenschaftliche) Anwender.

Die technischen, „internen“ Probleme der Google-Buchsuche stehen diesen genannten, bereits allgemein bekannten und breit diskutierten „externen“ Kritikpunkten in nichts nach, werden aber weit weniger ins Blickfeld der Diskussion gerückt. Im Folgenden wird es deshalb darum gehen, praxisorientiert die Inhalte und Suchmöglichkeiten des Online-Tools – einschließlich ihrer Grenzen – kritisch zu beleuchten.

Die Diskussion orientiert sich dabei vor allem an den Bedürfnissen der (germanistischen) Lexikologie.⁴ Zunächst werden die wichtigsten Suchoptionen der „erweiterten Buchsuche“⁵ (z.T. im Vergleich zu COSMAS II des IDS) beschrieben. Diese sind:

- „mit **allen** Wörtern“:
Google findet diejenigen Texte, in denen sämtliche Suchbegriffe unabhängig von der Reihenfolge ihrer Eingabe vorkommen (entspricht dem logischen Operator UND)
- „mit der **genauen Wortgruppe**“:
Google findet diejenigen Texte, in denen alle Suchbegriffe in der Reihenfolge ihrer Eingabe vorkommen (entspricht dem Wortabstandsoperator w/+1)
- „mit **irgendeinem** der Wörter“:
Google findet bis zu 32 verschiedene Suchbegriffe (entspricht dem logischen Operator ODER; diese Suchfunktion eignet sich besonders für Flexionsformen und Schreibvarianten)
- „**ohne** die Wörter“:
Google schließt alle diejenigen Texte vollständig aus dem Suchergebnis aus, in denen der Suchbegriff mindestens einmal vorkommt (entspricht dem logischen Operator NICHT)

Optionale Sucheinschränkungen

Die Google-Buchsuche bietet zusätzlich eine Reihe optionaler Sucheinschränkungen:

- Sprache
- Titel
- Autor
- Verlag
- Veröffentlichungsdatum
- ISBN

Die Nutzbarkeit all dieser Suchoptionen ist stark von der OCR-Qualität (s. dazu unten) abhängig. Was das Feld „**Sprache** Antwortseiten, geschrieben in“ leistet, muss unklar bleiben, es ist kein (funktionierender) Sprachfilter. Das Feld „**Autor** Bücher von diesem Autor ausgeben“ ist unbrauchbar, wenn zwischen einem Autor und einem Herausgeber unterschieden werden muss. Die Suchen nach Titel, Verlag und ISBN sind ebenso unsicher, zudem dürften sie von Lexikologen wohl nur in Sonderfällen genutzt werden.

Die Suchoption „**Veröffentlichungsdatum** Nach Büchern suchen, die veröffentlicht wurden zwischen“, die (durch die freie Eingabe zweier Jahreszahlen) die zeitliche Einschränkung der Suche ermöglicht, mag dagegen einer der wichtigsten Gründe für die Nutzung der Google Buchsuche durch den Lexikologen sein. Sie ermöglicht unter anderem:

- die Suche nach Früh- und „Erstbelegen“
- das gezielte Füllen von Beleglücken
- das Ersetzen von Wörterbuch-Buchungen durch „echte“ Belege
- das Ergänzen eines bislang nur in zu eng geschnittener Form (etwa auf einem Belegzettel) vorliegenden Belegtexts
- das Entdecken neuer (historischer) Sublemmata

Beispiel: Artikel „Gouverneur“ im DFWB

Unter dem Lemma „Gouverneur“ in Band VI des Deutschen Fremdwörterbuchs⁶ sind insgesamt 27 Sublemmata versammelt. Davon sind sechs nicht in den Korpora des IDS belegt: „Gubernatrix“, „Gubernierer“, „gubernieren“, „Gubernierung“, „Gouvernierung“ und „Gubernalismus“, hinzu tritt „gubernial“, das nur in zwei Zusammensetzungen und zudem sehr vereinzelt belegt ist.⁷ Mit Hilfe der Google-Buchsuche war es möglich, für alle diese Ausdrücke (z. T. lückenhafte) Belegstrecken zusammenzustellen.

Eine solche Suche führt zur Ausgabe einer Trefferliste, die durch copyrightbedingte Einschränkungen in sich heterogen ist:

- „vollständige Ansicht“
So gekennzeichnete Texte sind für alle Nutzer vollständig einseh- und benutzbar. Diese Texte können zudem in Form von pdf-Dateien downgeloadet (und danach etwa ausgedruckt) werden.
- „eingeschränkte Vorschau“
Auch diese Texte sind vollständig digitalisiert, es werden aber nur ausgewählte Seiten angezeigt. Dies kann zum einen durch Restriktionen bedingt sein, die der herausgebende Verlag Google auferlegt, zum Anderen mit Googles Umsetzung des Urheberrechts zusammenhängen: Die Buchsuche blockiert pauschal den Zugang zu Volltexten von Büchern, die nach 1864 erschienen sind (sowie häufig den Zugang zu noch älteren) für Nutzer außerhalb der USA. Dies betrifft auch nach deutschem Recht gemeinfreie Texte (deren Autoren also bereits länger als 70 Jahre tot sind).
- „Auszug“
Treffer, die so gekennzeichnet sind, sind noch stärker restringiert, Google gibt nur Minimalkontexte aus, die in den meisten Fällen nicht einmal ganze Sätze beinhalten. Diese Auszüge sind aus verschiedenen Gründen für die wissenschaftliche Arbeit ungeeignet (s. u.).

Daher kommt zwar Religion in der Geschichte immer und überall vor; aber wie steht es mit Gott? Kommt auch Gott legitimerweise in der Geschichte vor? Man kann feststellen, daß moderne Geschichtswissenschaften die Geschichte aus der Mittelalterzeit und mittelalterliche Sogenannter „Auszug“

- „keine Vorschau verfügbar“

Von solchen Texten ist nicht einmal ein Minimal-kontext einsehbar, sie sind lediglich bibliographisch erfasst (zur Qualität bibliographischer Daten in der Buchsuche s. u.). Treffer in solchen Texten müssen nötigenfalls per Hand nachgeschlagen werden.

Diese copyrightbedingten Einschränkungen können allerdings von vornherein in die Suche einbezogen werden. So kann man nach dem Suchwort im Gesamtbestand suchen, oder aber Texte ausschließen, die nur in Auszügen sowie eingeschränkt verfügbar sind. Eine zusätzliche Suchoption erlaubt die Durchsicht von Bibliothekskatalogen.

Die Funktion „Meine Bibliothek“

Das Vorhandensein eines eigenen Google-Kontos vorausgesetzt, kann der Nutzer eine virtuelle Sammlung für ihn interessanter Bücher als persönliches Teilkorpus des Google-Korpus selbstständig zusammenstellen und separat durchsuchen lassen. Dies kann das Auffinden qualitätvoller Belegtexte zwar erheblich erleichtern, allerdings bringt diese Funktion auch einige Nachteile mit sich, insbesondere dann, wenn das persönliche Teilkorpus einen gewissen Umfang überschreitet. So sind die ausgewählten Bücher nicht (z. B. alphabetisch, chronologisch, textsortenspezifisch oder nach Sachgebieten) sortierbar, der Zugriff ist rechnergebunden, d. h. immer vom gleichen Rechner aus möglich, und die Suchoptionen sind eingeschränkt.

Grenzen und Probleme

Zeitliche Begrenzungen

Eine zeitliche Untergrenze anzugeben, jenseits derer die Benutzung der Google-Buchsuche nicht mehr sinnvoll ist, ist schwierig. Erfahrungsgemäß kann man im 17. Jahrhundert eine solche Grenze ziehen. Ältere Texte (mhd., ahd., mnd. etc.) sind zwar durchaus vorhanden, aber natürlich nur in späteren Ausgaben, so dass eine Suche mit Hilfe der Option „Veröffentlichungsdatum“ sinnlos ist. Ältere Texte sucht man daher besser in anderen Textdatenbanken (z. B. MHD-BDB, Bibliotheca Augustana, Titus).

Technisch bedingte Probleme

Kennzeichnung des Suchwortes in der Fundstelle

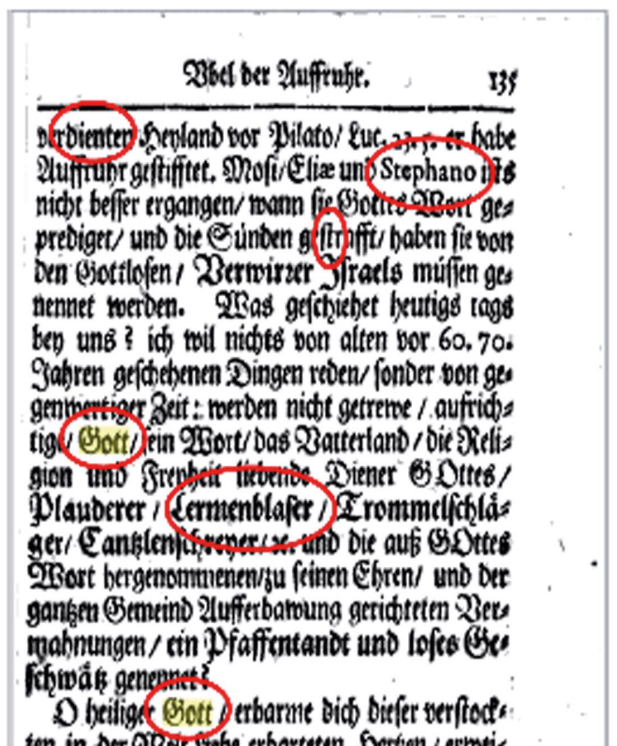
Das Suchwort wird im Fundstellentext gelb markiert. Nicht selten ist die Gelbmarkierung aber verschoben, so dass man das Wort auf der ausgegebenen Seite suchen muss. Manchmal ist es auf der ausgegebenen Seite gar nicht vorhanden.

OCR-Qualität

Die Bücher werden gescannt und durch OCR (Optical Character Recognition) als E-Texte in den Google-Index aufgenommen. Dabei entstehen die üblichen Fehler (z. B. wird das Schaft-s als f gelesen), wodurch es passieren kann, dass das gesuchte Wort, obwohl es eigentlich vorhanden ist, nicht gefunden wird (bzw. etwas als Fundstelle ausgegeben wird, das durch eine Fehlesung entstanden und gar kein wirklicher Treffer ist).

Google bietet zusätzlich zum Scan die Möglichkeit, die Ergebnisse der OCR in einer „Nur-Text-Fassung“ einzusehen. Textausschnitte (und ganze Texte) sind so per „cut & paste“ in eine Textverarbeitung übertragbar. Diese Nur-Text-Fassungen sind durch die Probleme der Texterkennung bedingt allerdings nur wenig verlässlich. Dies scheint sich erst seit Anfang 2008 zu bessern.

Der Vergleich eines Scans mit der entsprechenden Nur-Text-Fassung soll die typischen Probleme verdeutlichen (Suchwort: „Gott“):



Seitenabbildung des Original-Textes in Google.

verdienten Heyland vor Pilato/ Luc. 2. 3. er habe
Auffruhe gestiftet. Mosi/ Eliä und 8«pl«no ist
nicht besser ergangen/ wann sie Gottes Wort geprediget/
und die Sünden gestiftet/ haben sie von
den Gottlosen/ Verwirrer Israels müssen ge-
nennet werden. Was geschieht! heutig tags
bey uns? ich wil nichts von alten vor 60. 70.
Jahren geschehenen Dingen reden/ sonder von gegenwertiger
Zeit: werden nicht getreue / aufrichtige/
Gott/ sein Wort/ das Vatterland / die Religion
und Freyheit liebende Diener Gottes/
Plauderer / Lermenblaser/ Trommelschläger/
Canßenschreyer/ und die auß Gottes
Wort hergenommenen zu seinen Ehren/ und der
gangen Gemeind Aufferbawung gerichteten Ver-
mahnungen/ ein Pfaffentand und loses Ge-
schwätz genennet?
O heiliger Gott / erbarme dich dieser verstockten

OCR-Erkennung des obigen Ausschnitts.

Trefferanzahl

Die ausgegebenen Trefferanzahlen stimmen nicht. Eine Suche am 3. 12. 2008 nach „Vorkonstantinisches“ ergab angeblich 316 Fundstellen („Bücher 1 - 100 von 316 in Vorkonstantinisches“), aber schon die Anzahl der Trefferlisten (mit eingestellten 100 Treffern pro Seite) konnte mit 2 auf diese Angabe nicht passen. Die tatsächlich vorhandene Trefferzahl belief sich dann auf 122 – OCR-bedingte Fehler nicht eingerechnet.

Gleiches gilt für die Seitenzahlangaben in der Trefferliste und auf der Einzelergebnisseite, diese Angaben müssen immer anhand des Scans rückgeprüft werden. Dies allein ist ein Grund, die Auszüge nicht zu benutzen, da auf ihnen so gut wie nie Seitenzahlen zu sehen sind.

Probleme und Lösungsmöglichkeiten

Abschließend werden einige Lösungsvorschläge vorgestellt, die die praktische Arbeit mit der Google-Buchsuche erleichtern können:

Problem	Lösung
Über die Funktion „mit irgendeinem der Wörter“ findet Google bis zu 32 verschiedene Suchbegriffe, allerdings unterscheiden sich die so generierten Trefferlisten von denen, die sich ergeben, wenn man Google einzeln nach den verschiedenen Suchbegriffen suchen lässt.	Dieses Problem lässt sich nur dadurch lösen, dass man die verschiedenen Suchmöglichkeiten (einzeln und kumuliert) kombiniert.
Bei Suchwörtern, die Homographen in anderen Sprachen haben, ergibt sich oft ein störendes Übergewicht fremdsprachlicher Suchergebnisse.	<p>a) Eine praktikable Lösung ist der Ausschluss von Wörtern, die in der nicht erwünschten Sprache hochfrequent sind (wie „hic“, „the“ oder „leur“) im Feld „ohne die Wörter“. Dabei ergibt sich jedoch ein Folgeproblem: der Ausschluss bezieht sich auf den Gesamttext eines Buches, d. h., wenn sich in einem deutschen Buch ein einziges fremdsprachliches Zitat findet, in dem das ausgeschlossene Wort vorkommt, wird der gesamte Buchtext aus dem Suchergebnis aussortiert.</p> <p>b) Eine weitere Möglichkeit besteht in der Eingabe des Suchwortes als Teil eines Syntagmas (z. B. bei Substantiven: mit Artikel).</p>
Google scheint über den Rechnerstandort eine sprachspezifische Treffersortierung vorzunehmen, so dass (in diesem Fall) deutsche Textstellen voranstehen. Diese Sortierung funktioniert allerdings nur ungefähr.	Die einzige Lösung besteht darin, „per Hand“ alle Treffer in der Trefferliste durchzusehen, um „versteckte“ brauchbare Ergebnisse nicht zu übersehen.
Begrenzt man die Trefferliste mit Hilfe der Option „vollständige Ansicht“, werden oft weniger in vollständiger Ansicht vorhandene Bücher angezeigt, als tatsächlich verfügbar sind – in einigen Fällen gibt die Google Buchsuche auch an, es seien gar keine Bücher in vollständiger Ansicht vorhanden. Lässt man sich daraufhin sämtliche Treffer ausgeben, finden sich darunter oft einige vollständige Bücher (mehr).	Benötigt man mehr als einen Beleg in größerem Kontext, bleibt oft nur die Option, sich eine vollständige Trefferliste ausgeben zu lassen, und daraus die in vollständiger Ansicht vorhandenen Bücher „per Hand“ herauszusuchen.
In der Trefferliste wird nur ein(e) Treffer(-auswahl) pro Band angezeigt.	Mit der Funktion „dieses Buch durchsuchen“ auf der Einzeltrefferseite lassen sich weitere Treffer finden, allerdings generiert auch diese Suche keine vollständige Ergebnisliste. Eine vollständige Liste lässt sich allenfalls über die Websites solcher kooperierender Bibliotheken generieren, die ihre Digitalisate zusätzlich über ein eigenes System anbieten (z. B. die Universitätsbibliothek von Michigan).
Die Auszüge enthalten zu wenig Kontext bzw. keinen vollständigen Satz (was es u. a. schwierig macht, Zitate als solche zu erkennen).	<p>a) Eventuell ist auf einem Auszug aus einem anderen Exemplar desselben Buches mehr Kontext zu sehen oder das Buch ist in einer anderen Ausgabe vollständig einzusehen.</p> <p>b) Eine Suche nach der Wortfolge am Beginn oder am Ende des Auszugs liefert oft einen neuen Auszug mit dem gewünschten Kontext.</p> <p>Allerdings gilt, dass die Seitenangaben zu Auszügen äußerst unzuverlässig sind.</p>
Bereits gefundene Treffer sind manchmal später nicht mehr auffindbar.	Eine Möglichkeit besteht darin, den Rechner zu wechseln, auf anderen Rechnern wird der vermisste Textauschnitt oft wiedergefunden. Wenn dies nicht hilft, oder kein zweiter Rechner zur Verfügung steht, hilft oft nur ein weiterer Versuch einige Tage später. Eine Garantie dafür, dass man einen solchen „verlorenen“ Beleg wiederfindet, gibt es allerdings nicht.

Die bibliographischen Angaben der Google-Buchsuche stellen sich häufig als unzuverlässig heraus, sind unvollständig oder fehlen ganz.	Mit dem Tool „beliebte Passagen“ kann nach Zitierungen der gesuchten Stelle in anderen Büchern gesucht werden. Dort sind die gesuchten Daten dann eventuell (z. B. in einer Fußnote) angegeben. Bibliographische Angaben sind zudem grundsätzlich (anhand des KVK (Karlsruher virtueller Katalog) o. ä. Hilfsmittel) rückzuprüfen.
Datierungen sind häufig falsch, insbesondere bei Zeitschriften. Diese sind meist nach dem Erscheinungsjahr der ersten Nummer datiert, was oft nur zufällig bemerkt wird.	Die Vergrößerung des Scans mit dem Browser (eine Möglichkeit, die nicht alle Programme bieten) macht die Jahreszahl nur in wenigen Fällen lesbar. Zielführender ist es, mit Hilfe der Funktion „in diesem Band suchen“ nach Wörtern wie „Band“, „Jahrgang“, „Subskription“ (oder ihren fremdsprachlichen Pendanten) zu suchen. Eine weitere Möglichkeit besteht darin, nach dem Wort „Inhaltsverzeichnis“ zu suchen und den Text über einen Aufsatztitel zu datieren. Äußerst unbequem, aber oft die einzige Lösung ist es, im Suchfeld nacheinander wahrscheinliche Erscheinungsjahre einzugeben.
Die Auflage wird von Google nicht angegeben.	Dieses Problem ist nur bei vollständig verfügbaren Texten oder mittels Erscheinungsjahr unter Hinzuziehung bibliographischer Hilfsmittel lösbar.
Oft werden von Google falsche Seitenangaben gemacht.	Mit dem Tool „beliebte Passagen“ kann nach Zitierungen der gesuchten Stelle in anderen Büchern gesucht werden. Dort ist die gesuchte Seitenzahl dann eventuell (z. B. in einer Fußnote) angegeben.

Die Zusammenstellung dieser Probleme (und einiger Lösungen) zeigt, dass die Google-Buchsuche ein durchaus brauchbares Hilfsmittel der Lexikologie darstellt, dass jedoch von einem Gebrauch von Texten jenseits der in vollständiger Ansicht verfügbaren Digitalisate aufgrund der Unsicherheit der Angaben durch das System (derzeit noch) abzuraten ist.

Anmerkungen

- ¹ Der Text geht auf einen Vortrag zurück, den der Autor zusammen mit Herbert Schmidt am 28. Juli 2008 im IDS gehalten hat.
- ² Kasner, Edward/Newman, James: *Mathematics and the Imagination*. New York: Simon and Schuster 1940.
- ³ „Gemeinfrei“ bedeutet für Texte, die über die Google-Buchsuche bereitgestellt werden, dass in den USA alle vor 1923 erschienenen Texte vollständig einsehbar sind, für Nutzer außerhalb der USA zieht Google die Grenze aber unabhängig von der jeweiligen nationalen Rechtslage bereits im Jahr 1864.
- ⁴ Die folgenden Beobachtungen ergaben sich aus der täglichen Arbeit am Deutschen Fremdwörterbuch, einem lexikographischen Projekt, das den Kernbereich der geläu-

figen, in die deutsche Standardsprache der Gegenwart fest integrierten Fremdwörter und Fremdwortfamilien in ihrer historischen Entwicklung beschreibt und dokumentiert.

- ⁵ http://books.google.de/advanced_book_search (1. Dezember 2008).
- ⁶ Deutsches Fremdwörterbuch. Begonnen von Hans Schulz, fortgeführt von Otto Basler. 2. Auflage, völlig neu erarbeitet im Institut für Deutsche Sprache. Band 6: *Gag – Gynäkologie*. Von Gerhard Strauß, Dominik Brückner, Isolde Nortmeyer, Herbert Schmidt, Oda Vietze unter Mitarbeit von Heidrun Kämper. Berlin/New York: de Gruyter 2008, S. 438-452.
- ⁷ Dazu ist eine ganze Reihe von Sublemmata nur vereinzelt in den IDS-Korpora belegt („gouvernemental“, „Gouvernementalität“, „Gouvernementalist“, „gouvernementalistisch“, „Gouvernementalismus“ etc., eine Wortfamilie, die angeblich auf Michel Foucaults Begriff „gouvernementalité“ zurückgeht, durch die Google-Buchsuche aber bereits im 19. Jh. nachgewiesen werden kann).

Der Autor ist wissenschaftlicher Mitarbeiter am Institut für Deutsche Sprache in Mannheim.